The Alberta Journal of Educational Research *Vol. XLIX, No. 3, Fall 2003, 244-251*

Shizuhiko Nishisato
Ontario Institute for Studies in Education of the University of Toronto

Total Information in Multivariate Data from Dual Scaling Perspectives

It is an established matter that the total information in multivariate data is defined as the sum of eigenvalues of the variance-covariance matrix. In this article I challenge this time-honored tradition and look at another definition of the total information in data from a dual scaling perspective. This proposal is a step toward unifying the concept of information for both discrete and continuous variables.

Consider a continuous variable. The information of the variable is defined as its variance. Suppose we add another variable so that we can express the location of each respondent in a two-dimensional graph with two coordinates. Assuming that both variables are centered, the variance of data points in two dimensions is the variance of those distances of points from the origin, and it is not difficult to see that the variance is the sum of the variances of individual variables following the Pythagoras theorem. It is not difficult to see either that the same is true even when the two axes are rotated to the principal axes, in which case the sum of the variances becomes the sum of the eigenvalues. This discussion can be extended to any number of variables, hence the traditional definition of the total information in multivariate data.

Nishisato (2002), however, showed that the above definition is limited using the following example. Consider five standardized variables. Then the five eigenvalues under two extreme cases are:

1. Perfect correlation: $\lambda_1 = 5, \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0$
2. Perfect independence: $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 1$

In both cases, the sum of eigenvalues is five, which is the total information in the data. The objections to this traditional definition come from the common sense (a) that if all five variables are perfectly correlated, only one variable is needed to explain the data because the other four variables are totally redundant, and (b) that if all the variables are uncorrelated, one needs all of them to explain the data. His conclusion, therefore, is that the data set of perfectly correlated variables contains much less information than that of totally uncorrelated variables.

The above view was tied to research on dual scaling of discretized continuous variables (Eouanzoui, in press; Nishisato, 2000, 2002), which aims for a unified treatment of both discrete and continuous multivariate data.

Shizuhiko Nishisato is a professor emeritus in measurement and evaluation.

Dual Scaling

Principal component analysis (PCA) is carried out with continuous variables, say **Z**, using singular-value decomposition (SVD, Beltrami, 1873; Eckart & Young, 1936; Jordan, 1874; Schmidt, 1907).

$$\boldsymbol{Z} = \boldsymbol{Y}\Lambda\,\boldsymbol{X}' \text{ , or } z_{ij} = \sum_{k=1}^{K} \lambda_k\, y_{ik}\, x_{jk} \text{ ,} \tag{1}$$

where y_{ik} and x_{jk} are singular weights of component k of row i and column j respectively and λ_k is the singular value of component k.

When each continuous variable is categorized and a set of categorized variables is subjected to PCA, it is what we call dual scaling (DS). This small step in the procedure is a giant step in its implication for data analysis. Instead of considering cross-products of variables, it now becomes considering cross-products of functions of variables, a jump from the Hilbert space to the Sobolev space.

Consider multiple-choice data with several response options per question. The main object of DS (Nishisato, 1980, 1994) can be stated in many ways, one of which is to determine option weights so as to maximize the average inter-item correlation. The task is also called principal component analysis of categorical variables (Torgerson, 1958), multiple correspondence analysis (Benzécri et al., 1973; Greenacre, 1984; Lebart, Morineau, & Tabard, 1977), homogeneity analysis (Gifi, 1990), and many other names.

To simplify our discussion, let us consider three response options per question. Then the possible response patterns to a question are (1, 0, 0), (0, 1, 0), and (0, 0, 1), where 1 indicates the choice of the option and 0 a non-choice. Because three columns of the response patterns are mutually exclusive, we need a three-dimensional space for each variable, and responses to an item from subjects fall at one of these three points and nowhere else. When data are collected, therefore, the locations of the three coordinates can be determined by specifying the scaling unit. No matter what scaling unit one may choose, it is clear that an item with three response options yields only three distinct points, and they can be mapped in two-dimensional space. So long as these points are distinct, we need a plane (two axes) to accommodate the points of each item. If two multiple-choice items with three response options each are perfectly correlated, however, the two triangles converge into a single triangle, and the data can be mapped in two-dimensional space, no longer in four-dimensional space, nor in one-dimensional space as one might expect from two perfectly correlated continuous variables. This is a key difference between categorical and continuous variables.

Data in Multidimensional Space

Suppose variable p has coordinates $(x_{p1}, x_{p2}, \cdots, x_{pK})$ in K-dimensional space and the configuration is centered. In practice, however, we often express data in the space of dimensionality smaller than K, say g. Then both the squared distance from the origin to the coordinates in g-dimensional space ($g < K$) and the squared distance between variables p and q generally increase as the dimensionality g increases toward K. For example, the distance between points p and q as viewed in one-dimensional space would either remain the same or increase if we looked at them in two-dimensional space or three-dimensional space.

Any additional move of a point due to another dimension can never reduce the distance from the other point or the distance from the origin.

Thus we can now state that the triangle created by connecting the locations of the three response options of an item in two-dimensional space increases its size steadily as the dimensionality of the space increases. In the total space, therefore, the triangle of a three-category variable becomes the largest, the area of which can be regarded as proportional to the total information of the variable (Nishisato, 2002, 2003). In these papers Nishisato stated a remarkable aspect of its geometry as follows.

If the frequencies of the three options are equal, then the variable forms a regular triangle in the total space, that is, a regular simplex with the center at the origin. This regular simplex has the property that no matter how one may rotate the triangle, the contributions of the three vertices to the two axes are equal, that is, 50% each. Furthermore, he stated that even when the triangle is not regular (i.e., the three option frequencies are different) the contributions of the three vertices to two axes remain the same and equal to 50% no matter how the axes are rotated. One can prove this remarkable property quite easily by the structure of a 3-by-3 diagonal contingency table, the eigenvalues of which are $\lambda_1 = \lambda_2 = 1$, irrespective of the three frequencies in the main diagonal positions. This conclusion is contingent on the fact that we use the chi-square metric (Lebart, Morineau, & Warwick, 1984; Nishisato & Clavel, 2003) in DS. These observations lead to the conclusion that any variable with three response options requires two dimensions and that the contributions of the three points on the two dimensions are equal, and equal to 50%. When the number of response options increases to m_j, the variable in the total space forms the $(m_j - 1)$-dimensional polyhedron, of which each contribution to the total space is exactly $100/(m_j - 1)\%$.

Research on discretization of continuous variables for use by DS (Eouanzoui, in press; Nishisato, 2000, 2002, 2003) is motivated by the idea that the information contained in continuous variables is much more than what is assessed by the sum of the eigenvalues of the variance-covariance matrix. For example, principal component analysis of continuous variables captures only linear relations among variables, which is reflected on the sum of eigenvalues; but the data must contain nonlinear relations as well. If we consider continuous variables as categorical variables with as many categories as the number of distinct values and subject the data to DS, then we will be able to capture not only linear relations, but also nonlinear relations. The amount of total information is then the sum of the eigenvalues of the variance-covariance matrix plus alpha, and this alpha is likely to be much larger than the sum of the eigenvalues. In this case we can regard the volume of a polyhedron created by the number of categories (no longer three, hence no longer a triangle) as proportional to the total information of the variable.

If one is afraid of the phenomenon of overquantification, rest assured that it is an unwarranted fear. If two variables are distributed as bivariate normal, the correlation between the variable, say ρ^*_{jk}, is an upper bound of the product-moment correlation calculated from the corresponding discretized variables, say r^*_{jk}, that is,

$$\rho^*_{jk} \geq r^*_{jk} \ . \qquad (2)$$

Joint Distribution of Information

Consider the difference between PCA and DS once more. In terms of geometry, each continuous variable can be expressed as an axis along which all variates of the variable are distributed. To simplify our discussion we use the principal coordinates for the data space, that is, the principal hyperspace. When all the variables are perfectly correlated to one another, we need only the first principal axis to represent the data; when all n variables are totally uncorrelated, we need n axes to represent the data.

Consider the same extremes with categorical data with three categories (options) per variable. When all variables are perfectly correlated to one another, we need two axes (dimensions) to represent the common triangle; when all n variables are totally uncorrelated, the first variable occupies, for example, the first two dimensions, the second variable dimensions 3 and 4, the third variable dimensions 5 and 6, and so on, so that all n triangles are disjoint, that is, they have no overlapping parts, thus requiring $2n$ dimensions. Adopting the definition of information being proportional to the area created by connecting all vertices of the data points, we can state without ambiguity that the total information in multivariate data depends on the covariation (correlation) among the variables.

Contributions of Components to Total Information

To consider the amount of information in data, it is convenient to redefine the contribution of a component to the total information separately for discrete and continuous variables.

Discrete Case: Dual Scaling

DS determines option weights so as to maximize the average inter-item correlation. Its objective function for optimization is typically the correlation ratio η^2, which is equal to the mean of the squared item-total (component) correlation coefficients

$$\eta_k^2 = \sum_{j=1}^{n} \frac{r_{j\,t(k)}^2}{n} \quad . \tag{3}$$

The correlation ratio is the eigenvalue for DS. Assuming that the number of respondents is larger than the total number of options (say m) minus the number of questions n, the sum of all the eigenvalues, that is, the traditional total information in the data is given (Nishisato, 1994) by

$$\sum_{k=1}^{K} \eta_k^2 = \overline{m} - 1 \ , \tag{4}$$

where $K = m - n$ and $\overline{m}$ is the average number of options. It is also known (Nishisato, 1994) that the sum of squared item-total correlations of variable j over all K components is equal to the number of options minus 1, that is,

$$\sum_{k=1}^{K} r_{jk(k)}^2 = m_j - 1 \ , \tag{5}$$

where m_j is the number of response options of item j.

The sum of the eigenvalues depends on what objective function is used. Therefore, we would like to propose the use of a statistic that is independent of the objective function. Specifically, we define the total information as the sum of squared item-total correlations. In our case, for solution (component) k,

$$\sum_{j=1}^{n} r^2_{j\,t(k)} = \eta\ \eta^2_k = T^*(inf_k)\ , \tag{6}$$

and for the total space,

$$\sum_{j=1}^{n}\sum_{k=1}^{K} r^2_{j\,t(k)} = n(\overline{m} - 1) = T^*(inf)\ . \tag{7}$$

Let us look at three distributions of $r^2_{j\,t(k)}$ for three multiple-choice items with three options each: (a) when all items are perfectly correlated; (b) a general case; and (c) when all items are uncorrelated. We can see that the item total contribution of 2 is distributed differently in the three cases. (a) The three triangles merge completely when all inter-item correlations are perfect; (b) three triangles float in the six-dimensional space with different orientations; and (c) each of the triangles occupies two dimensions different from the others (Table 1).

Continuous Case: Principal Component Analysis

We typically use λ for the eigenvalue. Suppose we standardize the variables and consider principal component analysis of the correlation matrix. Then

$$\lambda = \sum_{j=1}^{n} r^2_{j\,t(k)} = T^*(inf_k)\ . \tag{8}$$

When we sum the squared item-total correlation of one variable over all possible components, we obtain

$$\sum_{k=1}^{K} r^2_{j\,k(k)} = 1\ , \tag{9}$$

where $K = n$. For the total space, we obtain

$$\sum_{j=1}^{n}\sum_{k=1}^{K} r^2_{j\,k(k)} = \sum_{k=1}^{n} \lambda_k = n = T^*(inf)\ . \tag{10}$$

Using three continuous variables, we can consider the distribution of $r^2_{j\,t(k)}$ under the same three cases as for the categorical cases discussed above. Note when these variables are perfectly correlated only one dimension is needed for the data (Table 2).

A New Measure of Total Information

We now move ahead with our goal of proposing a statistic for total information contained in multivariate data, continuous or discrete. Although it is not clearly stated, we should note that the traditional definition of total information is based on the case of independent variables. For example, when we have five multiple-choice items with three response options per item, we state that we need 10 dimensions to accommodate the data. But it is made clear above that

Table 1
Squared Item-Total Correlation: $r^2_{jt(k)}$
Categorical Variables with Three Options per Item and DS

Dimension (Component)	*1*	*2*	*3*	*4*	*5*	*6*	*Sum*
Perfect inter-item correlation							
Item 1	1	1	0	0	0	0	2
Item 2	1	1	0	0	0	0	2
Item 3	1	1	0	0	0	0	2
Any values of correlation							
Item 1	a	b	c	d	e	f	2
Item 2	g	h	i	j	k	m	2
Item 3	n	o	p	q	r	s	2
Zero inter-item correlation							
Item 1	1	1	0	0	0	0	2
Item 2	0	0	1	1	0	0	2
Item 3	0	0	0	0	1	1	2

this argument is based on the case that each variable occupies two-dimensional space independently of the other variables, hence 2 x 5 = 10 dimensions; if the variables are perfectly correlated to one another, we need only two dimensions. Thus we should note that the cloud of data points changes its volume depending on the inter-item correlations.

The statistic $T^*(inf)$ discussed above is independent of correlation among variables, and as such it contradicts the purpose of this article. We propose the following measure of total information for both discrete and continuous variables:

$$T(inf) = T^*(inf) - \sum_{i<j}^{n} r_{ij} + \sum_{i<j<k}^{n} r_{ijk} - \cdots + (-1)^{K-1} r_{123\ldots K} \,, \tag{11}$$

where

$$r_{123\cdots p} = \sum_{i=1}^{N} \frac{z_{1i}\, z_{2i}\, z_{3i} \cdots z_{pi}}{N} \tag{12}$$

z_{ji} is the standardized score of subject i on item j, and N is the number of subjects.

The new statistic can be interpreted as a concept corresponding to the union of sets in set theory and the joint entropy in information theory. As is clear from set theory, the union is the sum of the unique parts of single sets, unique parts of a pair of sets, and so on. In set theory and information theory, the traditional total information corresponds to the sum of sets and the sum of entropies of individual variables, respectively.

The new measure of information attains its minimum when all variables are perfectly correlated to one another and attains its maximum when all variables are totally uncorrelated. If continuous variables are subjected to PCA as is usually done, the analysis captures only linear relations. Then,

Table 2
Squared Item-Total Correlation: $r^2_{j(k)}$
Continuous Variables and Linear Analysis

Dimension (Component)	*1*	*2*	*3*	*Sum*
Perfect inter-item correlation				
Item 1	1	0	0	1
Item 2	1	0	0	1
Item 3	1	0	0	1
Any values of correlationl				
tem 1	a	b	c	1
Item 2	d	e	f	1
Item 3	g	h	i	1
Zero inter-item correlation				
Item 1	1	0	0	1
Item 2	0	1	0	1
Item 3	0	0	1	1

1. Continuous Variables for Linear Analysis

$$1 \leq T(inf) \leq n \tag{13}$$

When DS is carried out with categorical data or discretized continuous variables, then

2. Discrete Variables

$$n(M-1) \leq T(inf) \leq n(\sum_{j=1}^{n} m_j - 1), \tag{14}$$

where M is the smallest value of m_j, $j = 1,2,\cdots,n$ and m_j is the number of categories of variable j.

Discussion

It may look strange that n continuous variables have less information than n categorical variables. At the present moment this is so unless a nonlinear model is used for analysis of continuous variables. The above bounds for continuous variables apply only to linear analysis. As briefly mentioned above, however, the aim of the current research is to assess the total amount of information in data, including linear and nonlinear relations. Thus the proposal is to treat continuous variables as categorical variables with many categories per variable and use the formula for categorical data for assessment of total information. One aspect of the lower bound of the measure for categorical variables needs to be explained. The lower bound is attained when all variables are perfectly correlated. Then if variables j with m_j categories and variable k with m_k categories are perfectly correlated and if $m_j < m_k$, then there exist at most m_j distinct weights. Otherwise, the correlation of 1 cannot be obtained. Hence we arrive at the above lower bound.

The traditional definition of total information in multivariate data is the sum of eigenvalues of the variance-covariance matrix, which is known to be the sum of variances of individual variables. It is obvious that this definition is based on the extraordinary assumption that all variables are mutually inde-

pendent of one another. In the current study we redefined the information in terms of the sum of the squared item-component correlation coefficients and then adjusted the measure with respect to covariances. This idea can be interpreted as defining the total information in terms of the volume of a polyhedron created by the dataset.

In applying the new measure, it is important to consider how many components we should use and how high the higher-order correlation coefficients we should use, for the formula will be too demanding when the number of variables increases.

The current study will be continued to the next stage in which the new measure will be reformulated in terms of set theory and information theory, as well as being extended to the case of continuous variables with both linear and nonlinear inter-variable relations. For this last problem we must develop a practical method for discretizing continuous variables.

Acknowledgments

This study was supported by a research grant to the author from the Natural Sciences and Engineering Research Council of Canada.

References

Beltrami, E. (1873). Sulle funzioni bilineari. In G. Battagline & E. Fergola (Eds.), *Giornale di Mathematiche, 11*, 98-106.

Benzécri, J.-P. et al. (1973). *L'analyse des données: II. L'analyse des correspondances*. Paris: Dunod.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 211-218.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Eouanzoui, K. (in press). *On desensitizing data from interval to nominal measurement with minimum loss of information*. Unpublished doctoral thesis, University of Toronto.

Jordan, C. (1874). Mémoire due les formes bilinieares. *Journal de Mathématiques Pures at Appliquées, Deuxième Série, 19*, 35-54.

Lebart, L., Morineau, A., & Tabard, N. (1977). *Techniques de la description statistique: Méthodes et logiciels pour l'analyse de grands tableaux*. Paris: Dunod.

Lebart, L., Morineau, A., & Warwick, K.M. (1984). *Multivariate descriptive statistical analysis*. New York: Wiley.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, ON: University of Toronto Press.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, NJ: Erlbaum.

Nishisato, S. (2000). Data types and information: Beyond the current practice of data analysis. In R. Decker & W. Gaul (Eds.), *Classification and information processing at the turn of the millennium* (pp. 40-51). Heidelberg: Springer-Verlag.

Nishisato, S. (2002). Differences in data structures between continuous and categorical variables from dual scaling perspectives, and a suggestion for a unified mode of analysis. *Japanese Journal of Sensory Evaluation, 6*, 89-94 (in Japanese).

Nishisato, S. (2003). Geometric perspectives of dual scaling for assessment of information in data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. Meulman (Eds.), *Recent developments in psychometrics* (pp. 453-462). Tokyo: Springer-Verlag.

Nishisato, S., & Clavel, J.G. (2003). A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika, 30*, 87-98.

Schmidt, E. (1907). Zür Theorie der linearen und nichtlinearen Integralgleichungen. Erster Teil. Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebner. *Mathematische Annalen, 63*, 433-476.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: Wiley.